DOCUMENT RESUME

ED 190 625 TM 800 419

| | |
|---|---|
| AUTHOR | Anderson, Beverly L.: And Others |
| TITLE | Guidelines for Selecting Basic Skills and Life Skills Tests. |
| INSTITUTION | Northwest Regional Education Lab., Portland, Oreg. Clearinghouse for Applied Performance Testing. |
| SPONS AGENCY | National Inst. of Education (DHEW), Washington, D.C. |
| PUB DATE | Apr 80 |
| GRANT | OB-NIE-G-78-0206 |
| NOTE | 39p. |
| AVAILABLE FROM | Northwest Regional Education Laboratory, 710 S.W. Second Ave., Portland, OR 97204 ($4.25). |
| EDRS PRICE | MF01 Plus Postage. PC Not Available from EDRS. |
| DESCRIPTORS | *Achievement Tests: Admission Criteria: Basic Skills: Daily Living Skills: Educational Counseling: Educational Diagnosis: Educational Planning: Formative Evaluation: Guidelines: Student Certification: Student Placement: Summative Evaluation: *Testing: *Test Reliability: *Test Selection: *Test Validity |

ABSTRACT

Before selecting an achievement test, it is essential to determine its purpose. There are eight purposes, relating to three educational decision making contexts: (1) instructional managerial (diagnosis, course placement, career guidance): (2) screening (selection and certification) and (3) programmatic (survey assessment, formative evaluation, summative evaluation). Test publishers and developers should report score consistency (reliability), validity, and logistics of administration. High reliability coefficients (.90-.99) are needed for screening: placement and guidance require a .80-.99 range: a lower range .70-.99 is acceptable for diagnosis and programmatic decisions. Teachers can determine validity by linking test content with their own predetermined instructional objectives, or with other tests and course grades. Publisher reported validity clusters around .90. Teachers and subject experts are the best validity judges of instructional and screening tests: administrators and teachers should judge validity for programmatic decisions. Scores can be referenced to a norm group or to instructional objectives. Logistics include thoroughness of instructions for administration, security, scoring, and costs. A list of 102 currently available tests, including achievement batteries, and tests of reading, mathematics, life skills, and language arts is appended. A publishers' directory and bibliography of test review sources are also appended. (CP)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

ED190625

# Guidelines for Selecting Basic Skills and Life Skills Tests

"PERMISSION TO REPRODUCE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY
NWREL
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Northwest Regional Educational Laboratory

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

TM 800 419

2

# Guidelines for Selecting Basic Skills and Life Skills Tests

Beverly L. Anderson
Richard J. Stiggins
Suzanne B. Hiscox

Clearinghouse for Applied Performance Testing
NORTHWEST REGIONAL EDUCATIONAL LABORATORY
710 S.W. Second Avenue
Portland, OR 97204

JUN 9 1980

This publication was prepared under grant No. OB-NIE-G-78-0206 from the National Institute of Education (NIE), Department of Health, Education and Welfare, by the Clearinghouse for Applied Performance Testing (CAPT), and published by the Northwest Regional Educational Laboratory, a private, non-profit corporation.

The opinions expressed in this publication do not necessarily reflect the position of the National Institute of Education, and no official endorsement by the Institute should be inferred.

CAPT Director: Beverly Anderson

NIE Project Monitor: Judy Shoemaker

Advisory Board:
Richard Brickley
Thomas Corcoran
Connie Kravas
Christine McGuire

APRIL, 1980

# PREFACE

Given the important role tests play in education, it is crucial that test users understand the fundamental principles of proper test use. These guidelines present some of those principles, focusing specifically on the selection and purchase of published basic academic skill and life skills tests. However, though the guidelines focus specifically on tests of basic and life skills, the principles presented here can be applied to review, selection, and purchase of most *achievement tests* intended for use in educational settings. Aptitude tests—those intended to measure a student's capacity to learn—are not covered here.

The guidelines are intended for use by (1) public school district test coordinators and administrators with limited testing expertise, (2) testing specialists who assist district testing committees in test selection, (3) state department personnel responsible for test selection, and/or (4) testing and measurement instructors teaching introductory courses in educational measurement. Though the guidelines are presented as a narrative written by a teacher serving on a districtwide testing committee and are intended for use by elementary and secondary educators, the principles presented are equally useful to postsecondary educators, adult basic education personnel, and educators involved in other adult testing situations.

To supplement the guidelines and further assist educators' test review and selection, the Appendices contain extensive lists of currently available basic skills and life skills tests. Information is presented on test characteristics, publishers, and sources of additional, more detailed information. Although these lists are intended to be quite comprehensive, inclusiveness is not claimed. Readers are urged to consult the reference documents cited in the Appendix for more comprehensive listings.

Beverly Anderson
CAPT Director

# ACKNOWLEDGEMENTS

Several people contributed significantly to the preparation of these guidelines. The authors are indebted to Barbara Cole, Wayne Neuberger, and Joan Wright for their invaluable critiques of early drafts of the guidelines. Special thanks are also due to Susan Gabriel and Sharon Tamara, who compiled the extensive lists of available basic skills and life skills tests, and to Warren Schlegel for his excellent graphic work. And, finally, our thanks to Vicki Spandel for her excellent editorial assistance.

Beverly Anderson, Richard Stiggins, and
Suzanne Hiscox, Authors

# TABLE OF CONTENTS


PAGE

**TESTING: BECOMING A CRITICAL CONSUMER** . . . . . 1

**THE MANY REASONS FOR TESTING** . . . . . 2
- Using Tests to Help Manage Instruction . . . . . 3
- Using Tests to Screen Students . . . . . 5
- Using Tests for Programmatic Decisions . . . . . 5
- Interrelationships Among Purposes . . . . . 7

**CHARACTERISTICS OF GOOD TESTS** . . . . . 8
- Test Score Consistency . . . . . 8
- Test Validity . . . . . 12
- Validity and Test Purpose . . . . . 13
- Validity and Score Interpretation . . . . . 14
- The Logistics of Testing . . . . . 15

**REFERENCES FOR TEST SELECTION** . . . . . 18

**SUMMARY AND FOLLOWUP** . . . . . 21
**APPENDIX A:** Current Available Basic Skills and Life Skills Tests . . . . . 23
**APPENDIX B:** Publishers' Names, Addresses, and Telephone Numbers . . . . 31
**APPENDIX C:** Reference Materials Describing and Reviewing the Tests . . . . 33

# TESTING: BECOMING A CRITICAL CONSUMER

*As a fifth grade teacher, I had always felt each day held its own challenge. However, I was unprepared for the new dimension my career was to take on—all because of a short stop at the principal's office last April. Our principal, Mrs. Harris, asked if I would be on the District Test Planning Committee. To be honest, the suggestion left me cold. I insisted that I knew next to nothing about testing, had had minimal training on the topic and, in general, wasn't very interested. I said I thought I was a very good teacher and that I should be putting my emphasis on instruction. Harris was unmoved.*

*"That's the very reason I want you on the committee," she responded. "You are an excellent teacher. You know how to plan instruction to meet student needs and you have a very practical orientation to test use."*

*Who can resist such flattery? I became the tenth member of a District Test Planning committee that included four teachers, two guidance counselors, a building principal, the district testing coordinator, a member of the school board, and a volunteer parent. A week later I found myself in a two-day workshop led by a measurement specialist, Julie Swanson, from a nearby Regional Educational Laboratory. One workshop can't make you a test expert but this one surely helped us begin to thoughtfully analyze our testing needs. The District Superintendent and School Board wanted us to concentrate on basic skills testing—and to check out life skills testing too.*

*Before the workshop, I had called Julie to get a sneak preview of the agenda and to hear her views on testing. Our conversation went something like this.*

*I asked: "What's happening in achievement testing these days?"*

"Well," she responded, "we've all been hearing a great deal about the accountability movement in recent years. That, combined with recent instructional innovations and federally mandated program evaluations has increased the use of basic skills tests in schools. To keep pace, test publishers have diversified—by designing completely new tests, or by revising score interpretation and recommended score use procedures for existing tests.

"Test users, in turn, have had to become pretty sophisticated in testing. And there are a number of informational sources available: recent textbooks on educational measurement, existing test evaluation standards, and personal training or technical assistance. But none of these sources can offer pat answers to many of the questions you have to ask yourselves. For one thing, even leading measurement experts routinely disagree about which testing practices are 'proper.' To complicate matters, not everybody understands testing terminology—and sources aren't always helpful. At times it seems as if there are as many ideas about what a 'good' test is as there are test users. But the need to review, evaluate and select tests regularly goes on just the same."

*"What's the secret to good test use?" I asked.*

"It's more than mastering technical jargon," Julie answered. "In fact, it's more a matter of understanding how testing purposes and test characteristics interrelate. Unless you have a clear picture of why you are testing, the kinds of decisions to be influenced by test data, the instructional outcomes to be measured, and characteristics of the student population, and unless you have access to current information on tests, you cannot expect to choose a good test for your purposes."

*I was hoping to hear more about testing purposes and characteristics in the workshop, and asked Julie what the agenda would cover.*

"This will be the first of several planning workshops we'll have. This first session is what I call an awareness workshop," she told me. "We'll examine the relationships among common testing purposes and important test characteristics to help you learn to choose tests that serve your needs. The first workshop lays the foundation of knowledge you need before beginning to select tests. Later, we will have what I call local application workshops. At that time we'll get into the details of your situation. We'll explore your testing needs and priorities and then we'll select tests to meet your needs. After those tests are administered, we'll go into the actual use of the results. It's a year-long process and should be fun—I'm looking forward to it! I'll conduct the various workshops and provide all necessary technical assistance along the way."

*I was beginning to see my role on the District Testing Committee as more significant. At the same time, I*

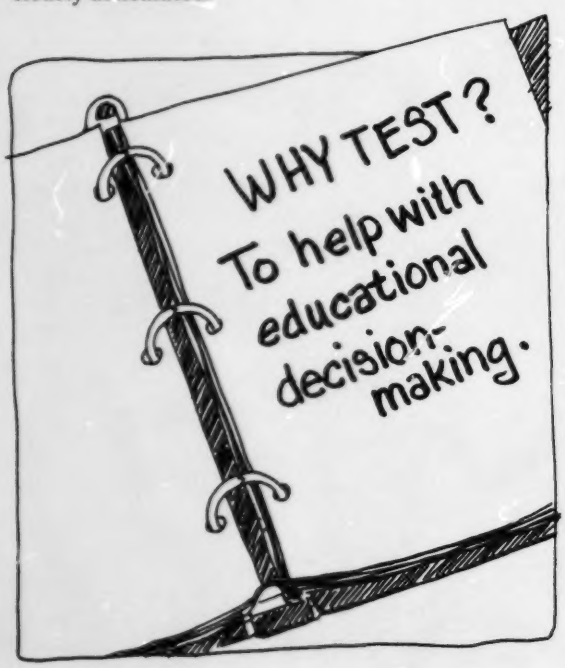
*was very relieved to know that Julie would be working with us throughout the year.*

*If you are involved in test selection, you may find it helpful to review this summary of the major points Julie covered in the first awareness workshop. She covered a lot of material in a short time—but in a non-technical way. She began by discussing the different ways you can use tests, and the characteristics of a "good" test. By showing us how to pick a good test for the intended purpose, Julie taught us to be good test consumers.*

## THE MANY REASONS FOR TESTING

*Julie opened the awareness workshop with some explanation of the kinds of decisions tests can be expected to help us make. She pointed out that she was about to outline the decisions as she thinks of them, though she admitted that others may organize the decisions in other ways.*

"Educational tests," she said, "have only one function: to facilitate educational decision making. A test should not be administered, therefore, until the decisions that rest on the results of that test have been clearly articulated.

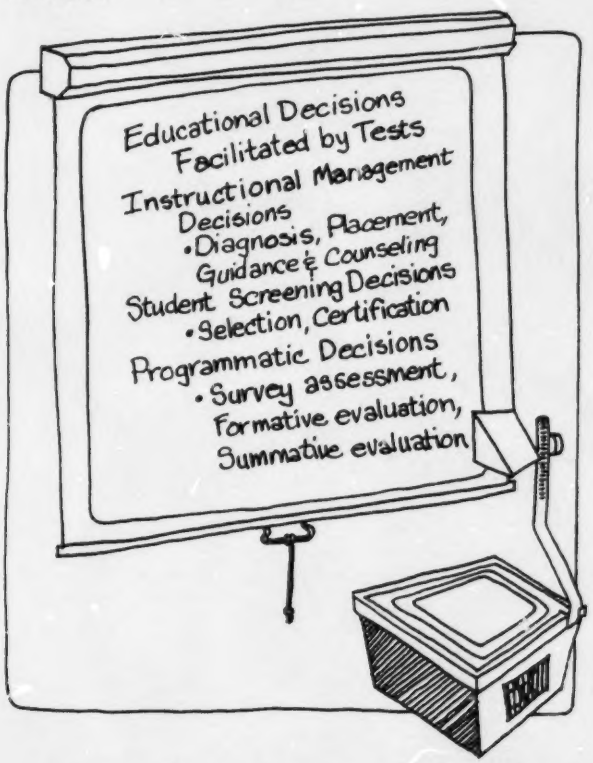



"Achievement tests are being used effectively in schools to help with three general types of educational decisions: instructional management, student screening, and programmatic decisions. Let's look at some of the specific decisions included in each of these categories.

"Basically, there are three instructional management decisions in which achievement tests can play a role: (1) *diagnosis* of learner strengths and weaknesses, (2) *placement* of students into courses, and (3) educational and vocational *guidance and planning*. In each case tests are used to link the student to the most appropriate level and type of instruction. But more about that in a moment.

"Achievement tests are also administered to screen students. For example, tests can help (1) *select* students for post-secondary programs (i.e., college or trade school), or (2) to *certify* students who have mastered certain minimum skills—as part of a certification program for high school graduation, for instance.

"Tests can also help in making decisions about educational programs. Specific instances include the use of tests for (1) *survey assessment*, (2) *formative* evaluation of instructional programs, and (3) *summative* program evaluation. Now, let's take each of these decision contexts apart and see if we can identify the key elements."


Educational Decisions Facilitated by Tests
Instructional Management Decisions
• Diagnosis, Placement, Guidance & Counseling
Student Screening Decisions
• Selection, Certification
Programmatic Decisions
• Survey assessment, Formative evaluation, Summative evaluation

*In the discussion that followed, Julie described each of these educational testing contexts in terms of the* specific decision *to be made, the* primary decision-maker, *and the* kind of information *that is often useful*

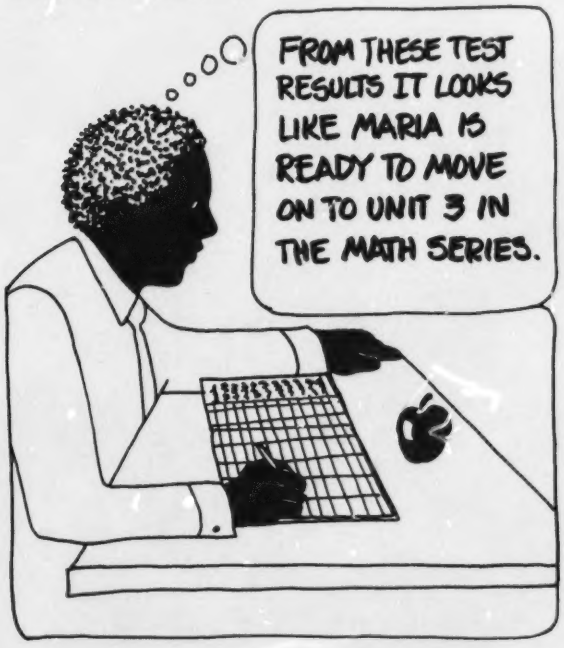
*in making the decision. At this point, I began to realize how we might counter teachers' apathy about the testing program by testing for a reason—not just because "we've always done it that way." While Julie was talking about the reasons for testing, she referred to the chart that appears here as Table 1.*

## Using Tests to Help Manage Instruction

*We teachers on the committee were eager to hear about instructional management. This was a somewhat familiar topic—one with which we felt at ease. Julie did a nice job of explaining the three ways tests are used to help link students to instruction. Generally, her discussion went like this:*

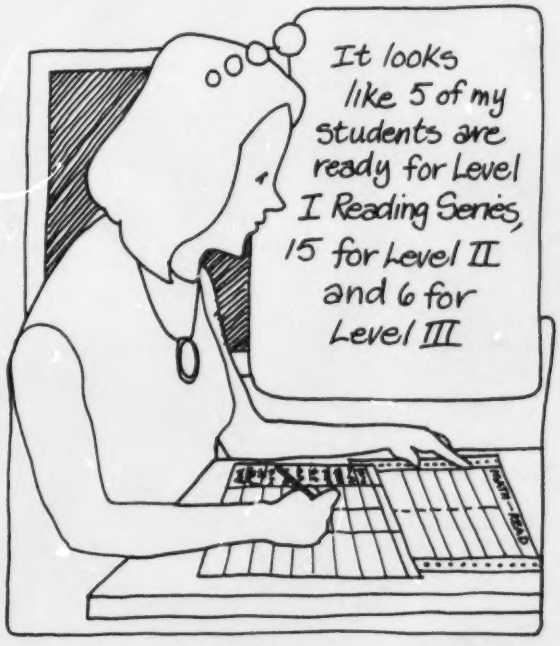
DIAGNOSIS Perhaps the most frequent reason for testing is to diagnose the strengths and weaknesses of individual students. In this context, the primary decision maker is the teacher, though the students themselves may also play a role in the decision. Teachers often use tests and other performance indicators to decide an individual student's current level of development so as to select the next most appropriate instructional unit for that student. Tests that are useful in this decision are those that reveal precisely what skills and knowledge the student has or has not mastered.

**Using Tests for Diagnosis**




PLACEMENT If diagnosis has as its goal the linking of students to individual instructional units within a course, placement can be thought of as linking the student to the next largest instructional element—the course or class. In this case, the primary decision makers are administrators and/or teachers (guidance counselors may also be involved) who have planned courses and need to ensure that each student enters the most appropriate type and level of course. For example, a math test might be used to place students at the appropriate level of math instruction in a high school math course sequence. Most useful here are tests that indicate students' level of math development so that students will not be assigned to courses in which the instruction is too advanced or too elementary for their skill level. Placement tests usually cover a broader range of knowledge and skills than diagnostic tests and are only used once or twice a year while diagnostic tests are used continuously. Grades and courses completed are also often considered in placement decisions.

**Using Tests for Placement**




GUIDANCE Diagnosis matches the student to an instructional unit, placement to a course, and guidance to an entire program of study. In this case, the primary decision makers are students and their parents, with assistance from guidance counselors. Each student must weigh the chances of success and satisfaction in various educational and vocational programs, then decide which to pursue. Such decisions are typically made in junior and senior high school when career planning becomes a prime concern. Tests of greatest value in guidance are those that cover broad academic areas and tell the students where they stand relative to other students. Test scores should also tell

Table 1: **Summary of Various Purposes for Testing**

| Assessment Context | Types of Decisions | Type of Data Needed | DECISION MAKERS | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | Students | Parents | Teachers | Administrators | Counselors | Public |
| **Instructional Management Decisions** | | | | | | | | |
| Diagnosis | Decide students' strengths and weaknesses on specific skills | Individual student data on level of development of specific skills | ✓ | | ✓ | | | |
| Placement | Place student into next most appropriate level of instruction | Scores that place students on relevant knowledge or skill continuum | | | ✓ | ✓ | ✓ | |
| Guidance & Counseling | Decide probability of success and satisfaction in given program of educational or vocational development | Data reflecting level of educational development of individual student relative to other students | ✓ | ✓ | | | ✓ | |
| **Student Screening Decisions** | | | | | | | | |
| Selection | Decide which students to be selected into or out of a program | Data that rank order individual students on relevant knowledge or skill scale | | | ✓ | ✓ | ✓ | |
| Certification | Determine mastery or non-mastery of specified body of knowledge or set of skills | Data reflecting individual student mastery of specified body of knowledge or set of skills | | | ✓ | ✓ | | ✓ |
| **Programmatic Decisions** | | | | | | | | |
| Survey Assessment | Make educational policy decisions; determine educational development of student group | Group achievement data gathered cyclically to show trends | | | | ✓ | | ✓ |
| Formative Evaluation | Decide program components in need of modification | Interim and final program outcomes attained and not attained by participating students considered as a group | | | ✓ | ✓ | | |
| Summative Evaluation | Determine if program is to be adopted, expanded, or discontinued | Program outcomes attained and not attained by participating students considered as a group | | | ✓ | ✓ | | ✓ |

students where their strengths and weaknesses lie so they can consider this information in choosing. Of course, test scores alone should never form the basis for any guidance decision. The student's academic record, interests and aspirations all merit consideration.

**Using Tests for Guidance and Counseling**

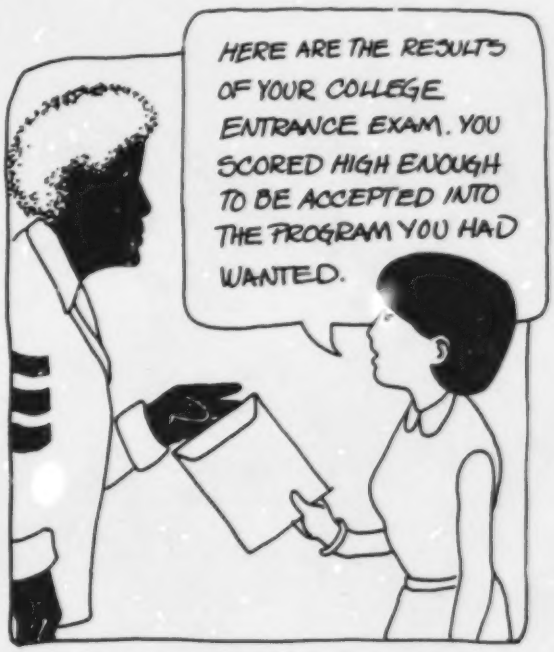
FROM THESE READING & MATH TEST RESULTS & WHAT YOU HAVE TOLD ME ABOUT YOUR INTERESTS, IT SEEMS LIKE YOU WOULD BE WISE TO CONSIDER GOING TO COLLEGE.

## Using Tests to Screen Students

*Since many of our students are interested in college, we were eager to hear what Julie had to say about uses of tests for selection. In addition, our district—like many others—is in the midst of a big debate on minimum competency testing and tracking students at the high school level. So our school board member really perked up when Julie began to discuss screening decisions. A summary of her comments follows:*

SELECTION The terms selection and placement are not always clearly differentiated. Placement, as described above, deals with the matching of all students tested to their appropriate level of instruction. This is an instructional management decision. The term selection is used here to refer to those decisions where a limited number of spaces are available in an educational program and entrants are screened for admission based on which are the most likely to benefit. The key decision makers are teachers and administrators of the selecting program. A test used for selection purposes (1) focuses on those skills and knowledge which are essential for future success, and (2) rank orders students in terms of relevant skills and knowledge to identify which students are most likely to succeed. College entrance or selection into a particular course (for example, airline pilot training) are prime examples. Generally, test scores are not the sole basis for selection decisions. Previous academic record is also considered.

**Using Tests for Selection**




CERTIFICATION Tests often play an important role in certifying those who have attained a specific minimum acceptable level of educational development. Consider just three examples of certification involving different decision makers: First, a teacher might use a test to certify mastery of beginning verbal skills to decide whether to promote a given student. Second, a school district administrator acting in the name of the board of education might use a test to certify attainment of minimally acceptable skills for receipt of a high school diploma. And third, members of a certain technical profession might use a test to certify competence to practice that profession. In each case, examinees must pass the test to be certified as minimally competent. Therefore, the test must focus specifically on clearly stated minimal competencies.

## Using Tests for Programmatic Decisions

*The discussion of programmatic decisions was a highlight of the workshop for me. I was particularly*

**Using Tests for Certification: The superintendent addressing the school board.**

"It's gratifying to report that 96% of our district students passed the high school certification test. The other 4% will receive remedial help & will have a chance to take the test again."

*interested in evaluating our school's reading program that year and had no idea what tests to use. We were getting pressure to use the same test the district had been using for several years, but I didn't know if that was right. I spent a lot of time talking to Julie after the workshop about this issue. She helped me clarify the three contexts in which testing can be used to make programmatic decisions: survey assessment, formative evaluation and summative evaluation. Here are my notes on her remarks.*

SURVEY ASSESSMENT One of the most established uses of testing in education is to survey student achievement and explore trends over time to make programmatic planning decisions. This kind of testing is usually designed to raise questions rather than answer them. For example, why are math scores gradually declining in the district (state or nation)? Or, why are reading scores of fourth graders consistently below national averages while those in other grades are above average? The test data are used to make decisions about what aspects of the education system need to be more thoroughly investigated to determine possible reasons for unsatisfactory performance. For this purpose, achievement test scores—sometimes gathered from random samples of students—are generally gathered on an annual basis, then averaged across the entire school, district or state to show the general level of student development. Test scores are frequently compared from year to year to show trends. This information becomes a basis for setting educational policy and allocating resources. Typically educational administrators are the key decision makers, but they must justify their decisions to the taxpaying public. In essence, tests covering broad content and skill areas are used to conduct an overall assessment of an educational program.

**Using Tests for Survey Assessment: State Department of Education spokesperson addresses a legislative committee.**

THE STATEWIDE ASSESSMENT RESULTS SHOW THAT OUR STUDENTS ARE WEAK IN WRITING SKILLS. I RECOMMEND ADDITIONAL FUNDS BE ALLOCATED TO DISTRICTS FOR WRITING INSTRUCTION.

FORMATIVE EVALUATION In formative evaluation, the goal is to determine which instructional units or features of a specific educational program (e.g., in remedial reading) are effective as they stand and which are in need of revision. The decision makers are program teachers and administrators. Tests covering the specific interim and long-term outcomes of the program are administered during the program's operation to help shape the program during its formative stages.

SUMMATIVE EVALUATION Summative evaluation seeks to reveal a program's overall merit, thus suggesting whether that program should be continued or terminated. Tests designed to assess students' performance on final learning outcomes of a program are typically a key part of such an evaluation. Teachers, program, building or district administrators, and the public (including the board of education) may be involved in summative evaluation decisions. Tests may be given prior to as well as following instruction, with retention testing after a given interval.

Achievement testing as a part of formative and summative evaluation is distinguished from survey assessment in that the former is part of a broader data collection effort. In particular, program evalua-

**Using Tests for Formative Evaluation: The staff of a special math project discuss their efforts**

THE RESULTS OF THE TEST WE GAVE AFTER 6 WEEKS OF USING THE NEW MATH SERIES INDICATE THE STUDENTS ARE HAVING MORE TROUBLE WITH FRACTIONS THAN WE EXPECTED. LET'S DISCUSS WHAT CHANGES WE MAY WANT TO MAKE IN THE PROGRAM.

**Using Tests for Summative Evaluation: The superintendent addresses a group of building principals.**

THE RESULTS OF THE READING TEST SHOW THAT STUDENTS IN THE NEW READING SERIES LEARNED MORE THAN THOSE IN THE OLD SERIES. ALSO TEACHERS LIKED THE NEW SERIES BETTER. I RECOMMEND ADOPTION OF THE NEW SERIES FOR ALL SCHOOLS IN THE DISTRICT.

tion data may include interviews with teachers or other school personnel or completed questionnaires. These data are used in conjunction with the test results to make evaluative decisions about programs. Survey assessment does not necessarily provide information on educational processes to assist in interpretation; rather, it is likely to indicate areas of educational deficiency that call for further investigation. For example, a random sample of fourth grade students across a state may suggest that students are weak in reading comprehension. Such results may prompt the state legislature to fund an evaluation of selected reading programs to determine the reason for these results.

## Interrelationships Among Purposes

*We were a bit overwhelmed by all these testing purposes. It sounded like we would be testing all the time. But Julie assured us that once we got all of the purposes straight in our minds, and could see how they interrelate, we could actually reduce our testing by using some tests to serve several purposes. That would be a big improvement over our present haphazard approach, I thought. Julie expanded her point with several examples:*

"First," she began, "you may have noted that the three programmatic decision contexts (survey assessment, formative and summative evaluation) don't require test scores for every individual student. Average test scores for the group will often suffice. Remember two things: First, when group performance data are needed, careful random sampling yields performance information every bit as useful as the data obtained by testing every pupil—at a fraction of the cost. Second, individual student scores can be averaged to produce group scores. This means that individual selection, certification, guidance and diagnostic test results can often be summarized for use in evaluation or survey assessment. A word of caution, though. Averaging is possible only when individuals included in the summary are truly representative of the total group. If every student in the group is tested, of course, this condition is clearly satisfied. But if only some students are tested, it is inappropriate to summarize individual data for a group decision unless you are certain the sample is representative.

"Second, remember that in instructional management and programmatic decision making, more specific scores can be collected over time and summarized to help with more general decisions. For example, diagnostic scores collected over time can be useful in placing a student in the next most appropriate course. And placement scores that rank order students can help each student select an appropriate educational and vocational course. Similarly, tests

such as college entrance exams often serve double duty as guidance and counseling tools. At the college level, selection tests can also be used in course placement.

"To illustrate further, formative evaluation information gathered and summarized over time can yield useful information for summative evaluation decisions, and summative evaluation data covering a broad enough program can often serve long range planning needs. Some careful planning and test score recordkeeping can economize efforts. Haphazard planning and poor recordkeeping result in inefficient, expensive testing. It's that simple."

"The key is to determine what decision must be made and to understand what tests will yield scores useful in making those decisions. Once this understanding is achieved, the trick is to get test scores to the right people at the best time. Some thoughtful planning can accomplish this.

## CHARACTERISTICS OF GOOD TESTS

*We spent the afternoon discussing characteristics of a good test. Small-group work kept the session lively. Julie started the discussion like this:*

"Whether choosing a test for a single purpose or multiple purposes, as a consumer, you should know how to judge the characteristics of a good test. Intelligent test purchase means knowing what you want a test to do for you and being able to pick the test that can do it. We have already talked about the kinds of decisions you might want to make. Now let's deal with the standards you, as a consumer, should use to determine whether a test will serve your purposes.

"Tests generally cover only a *sample* of the content or skills a test developer intends to measure. In most cases, it would be too time consuming and expensive to test all knowledge in a given area. But if the test is well constructed, the test user can generalize from the sample to a larger body of knowledge, thereby estimating an examinee's level of development. The test developer's goal is to construct tests that (1) estimate a student's level of development consistently, (2) estimate level of development accurately and (3) are relatively cost effective—in terms of both time and money. The quality of a test is judged by these three factors."

*Julie described each of these test quality factors in some detail from three perspectives: (1) the definition(s) of the factor and reasons for its importance; (2) the relationship between the factor and various testing purposes and (3) relevant information that should be provided by the test publisher.*

*Before providing us with step-by-step guidelines for evaluating a given test, Julie commented on the relative importance of these three test quality factors:*

"You should not infer from the order in which we will discuss these factors that the first is more important than the last. In fact, they are all important, and what you need to do is to find the best combination for your purposes. Obviously, a prime determining factor in test selection and purchase is the amount of money you have to spend. We will cover cost considerations as our third factor. But as you will see, cost is by no means the only consideration. It does little good to use a test that does not fit your purpose just because it is inexpensive. You might just as well use no test at all. As a matter of fact, that would be better than using the wrong test. Similarly, it does little good and can do great harm to use a test that yields inconsistent scores just because it's inexpensive. If evidence of score consistency (explained below) is not available for a test, don't buy it or use it—use some other performance indicator. The point I am making is that you need to consider all factors and find the best combination—observing certain minimum standards for each factor. This point will become more clear as we discuss each factor."

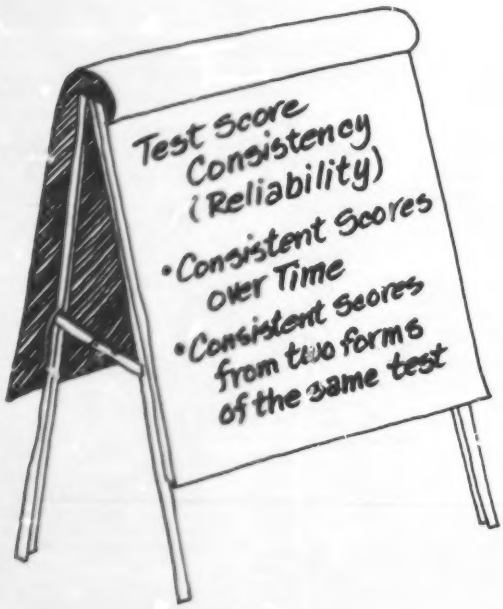
### Test Score Consistency

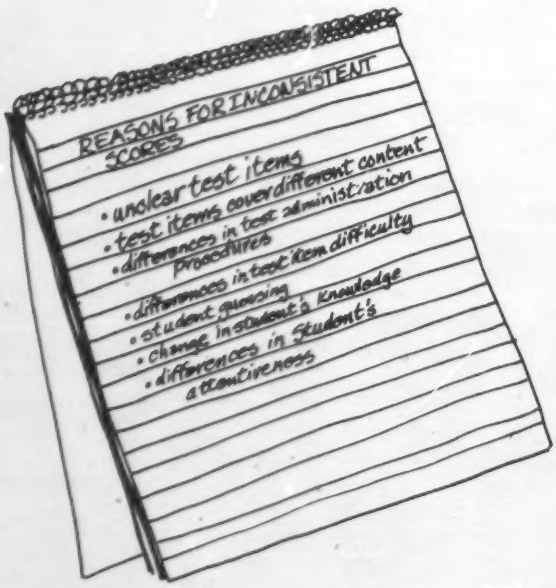
"Tests used in educational decision making must yield reasonably consistent scores. Otherwise, one cannot accurately estimate an examinee's level of development. Measurement specialists have several ways to define and evaluate consistency."

DEFINITIONS OF CONSISTENCY "Test score consistency can be defined in several ways. For example, tests should yield scores that are consistent over time. If the same test is administered to the same examinees twice a few weeks apart, with no instruction during the interim, the first and second scores should be about the same for each examinee. If they differ significantly for large numbers of examinees, the examiner will have no idea which score (if either) to rely on.

"Test score consistency can also have another meaning. Two sets of test items intended to measure the same knowledge or skills, and administered to the same examinee at the same time, should yield very similar scores.

"To illustrate, suppose two ostensibly equivalent forms of a test are constructed to measure certain skills in chemistry. Yet the scores achieved by any given chemistry student on the two tests are vastly different. It might be that the test items are not clear or one set of items is more difficult than another. Similarly, it may be that the tests were not administered correctly, or they just didn't cover the same material. Here again, the examiner would not know which





REASONS FOR INCONSISTENT SCORES
• unclear test items
• test items cover different content
• differences in test administration procedures
• differences in test item difficulty
• student guessing
• change in student's knowledge
• differences in student's attentiveness

score to rely on. Apparently, some factor(s) other than examinees' level of development are influencing performance, rendering the scores unreliable.

"Test publishers use both meanings of consistency in demonstrating to users that their test scores are reliable. They will have field tested their test during its development and will present evidence of score consistency resulting from that field trial. For example, the technical manual and/or test user's guide may show that scores are stable over time. This is called *test/retest reliability*. Or they may show that two equivalent test forms yield equivalent scores. This equivalence generally takes one of two forms. Either (1) the publisher will develop two completely independent and parallel (in terms of content and difficulty) forms of the test and will show that a given examinee will score at the same level on each test form, or (2) the publisher will divide the items in a single test randomly into two parts—all even-numbered items and all odd-numbered items, for instance—and show that an examinee's performance on the two parts is the same. The first form of equivalence is often called *parallel forms reliability*, while the second form of equivalence is termed *internal consistency reliability*. Both are forms of score consistency."

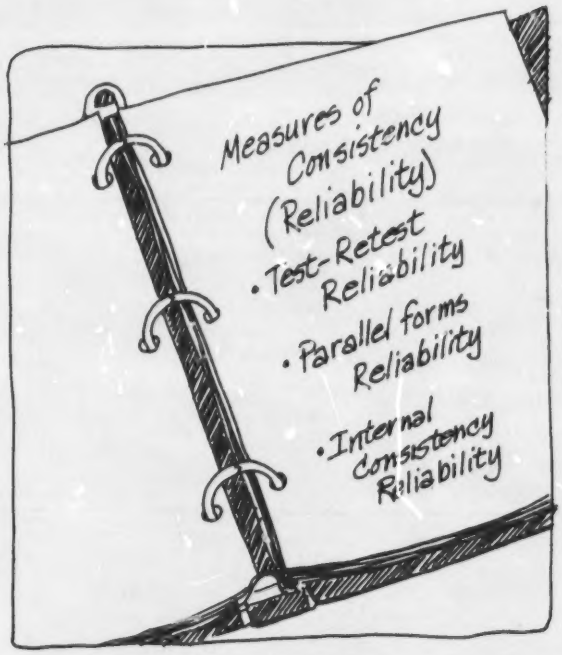
INDICES OF RELIABILITY "Measurement specialists have devised indicators of reliability that reveal the extent to which scores vary over time, across forms, and across parts of a given test. We are not going to go into a great deal of detail about the indicators, but you will need to recognize them in order to evaluate tests. When test publishers refer to *test/retest reliability*, they will be talking about the correlation between the scores achieved on the first test and the scores achieved on a second administration of the test. If scores tend to be stable over time, test correlation will approach 1.00. If it is below .70, depending on how you plan to use the test, there is reason to be suspicious about the consistency of the scores.

"When test publishers refer to *parallel forms reliability*, they will cite the coefficient of equivalence or the correlation between scores achieved by examinees on two ostensibly equivalent forms of the same test. If the forms are indeed equivalent, the correlation will approach 1.00. If it is at or below .70, depending on the intended use of the test, be suspicious.

"And finally, when test publishers refer to *internal consistency*, they are referring in a general way to the correlation between scores achieved by examinees on two random halves of a given test. Labels will include split-half reliability, KR-20, and/or coefficient alpha. Again, if the test tends to assess the same content and skills throughout (i.e., is internally consistent) the correlation will approach 1.00. The .70 lower limit also applies here, depending on purpose.

"Understanding score reliability helps the interested educator evaluate test quality based on the information presented by the test publisher. It is not possible within the scope of this workshop to comprehensively discuss the origin and meaning of each of these reliability indicators. Those of you who are interested in learning more about reliability are urged to seek the instruction of a qualified measurement specialist and to study a standard educational testing textbook."

RELIABILITY STANDARDS AS A FUNCTION OF TEST PURPOSE "All test publishers aspire to perfect score consistency. For various reasons, however, they virtually never achieve that goal. Factors over which no one has control will render all test scores slightly inconsis-




tent. For example, the way in which a test is administered may vary slightly from time to time. Examinees may be more alert at one time or another. Or they may guess correctly more often on one test than on another. The test developer's goal is to minimize these extraneous factors by carefully constructing tests and standardizing test administration procedures.

"The test user should select a test with sufficient score consistency to fulfill an intended purpose. Some of the eight purposes we talked of earlier demand more score consistency than others. Generally, the level of consistency required depends on (1) how serious an incorrect decision would be, (2) the number of examinees *directly and individually* affected by that decision, and (3) the extent to which the test data contribute to the decision. Here's an example: Suppose a test were given to a group of examinees, only a few of whom would be selected to receive a full four-year college scholarship. If the test did not yield near-perfect score consistency (reliability), then another administration or use of a parallel test might result in selection of different scholarship winners. If this were the case, then nonrecipients would be quite justified in challenging the selection procedure. Their success would be determined by the test used rather than by their own ability. In this example, each examinee is directly affected, the cost of error is high and the test is the sole basis for the decision. Therefore, high score reliability is a must.

"Other testing situations involve similar contingencies. If scores are used to certify competence, place individuals at different levels, make educational or career decisions, or manage instruction in individualized settings, very high consistency is necessary.

"Though score consistency is always important, *slightly* lower consistency can be tolerated within group decisions, because averaging student scores to obtain a group score can compensate somewhat for lower individual student score reliability. Also, the educational policy or program evaluation decisions based on group scores generally do not have the same negative impact on individual students as would, say, an incorrect scholarship decision. Additionally, other information besides test scores usually is weighted heavily when making programmatic decisions."

Some relatively specific guidelines for determining appropriate levels of reliability are conveyed in this summary chart:

Highest reliability needed: (r = .90-.99)
- Selection
- Certification

Next highest reliability needed (r = .80-.99)
- Placement
- Guidance

Lower reliability allowed (r = .70-.99)
- Survey Assessment
- Formative Evaluation
- Summative Evaluation
- Diagnosis

*Then Julie provided us with a neat summary (Table 2) of the relationship between reliability considerations and test purpose.*

*Julie reminded us that publishers must adhere to professional standards in providing information about their tests. Those standards include requirements for reporting on reliability. She shared a few guidelines from a booklet called* Standards for Educational and Psychological Tests, *published by the American Psychological Association, just to give us the flavor of the standards. She urged everyone to obtain a copy and read the complete set of standards.*

*Here are a few of the guidelines she mentioned:*

1. The test manual or research report should present adequate evidence of reliability to permit the reader to judge whether scores are sufficiently dependable for intended uses.
2. Reliability coefficients should be provided for every score, subscore, or combination of scores (such as sum, difference, or quotient) recommended by the test manual.
3. The procedures and samples used to determine reliability coefficients should be described sufficiently to permit users to judge the applicability of the data reported to the individuals or groups with which they are concerned.
4. If two or more test forms are published for use with the same examinees, information . . . should be reported in the test manual . . . on correlations among scores. If necessary evidence is not provided, the test manual should warn the reader against assuming score equivalence.

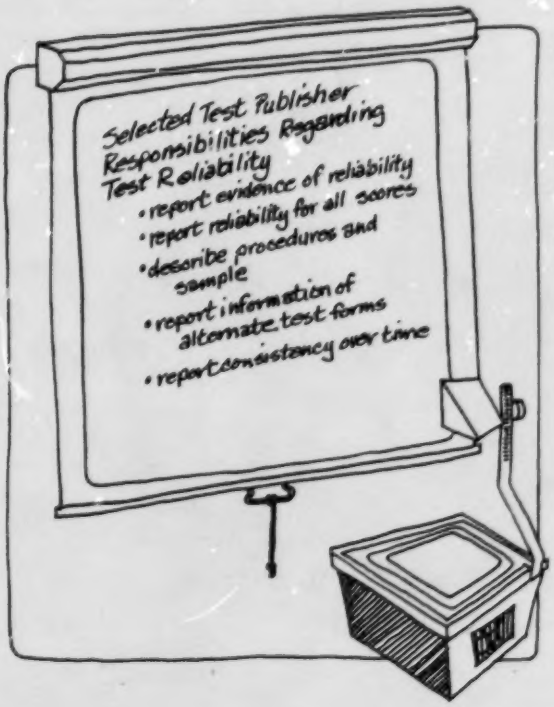
## Table 2: Key Elements in Evaluating Test Score
## RELIABILITY
### as a Function of Test Purpose

| Important Questions to be Addressed | INSTRUCTIONAL MANAGEMENT | | | SCREENING | | PROGRAMMATIC DECISIONS | | |
|---|---|---|---|---|---|---|---|---|
| | Diagnosis | Placement | Guidance/ Counseling | Selection | Certification | Survey Assessment | Formative Evaluation | Summative Evaluation |
| **General Question:** Can we rely on the scores to reflect learner traits consistently? | ☆ | ☆ | ☆ | ☆ | ☆ | ☆ | ☆ | ☆ |
| The published test manual and your test planning should address the following **more specific** questions: | | | | | | | | |
| 1. Is reliability estimated for every score included in the examinee performance profile resulting from the test? | ☆ | ☆ | ☆ | ☆ | ☆ | ☆ | ☆ | ☆ |
| 2. Is there evidence of score consistency over time? | ☆ | ☆ | ☆ | ☆ | ☆ | √ Desirable if available; often cost of gathering test/retest data prohibitive. | √ | √ |
| 3. If there are parallel forms of the test available, has their score equivalence been established? | ☆ | ☆ | ☆ | ☆ | ☆ | ☆ | ☆ | ☆ |
| 4. Is there evidence of internal consistency reliability?† | ☆ | ☆ | ☆ | ☆ | ☆ | ☆ | ☆ | ☆ |
| 5. Is there reason to suspect scores are based on very few items? If yes, is the reliability of the very short test addressed? | ☆ | √ Generally call for norm referenced tests where length is typically not an issue. | | √ | ☆ | √ Important with criterion referenced tests; with norm referenced tests, generally not an issue. | ☆ | ☆ |
| 6. Do the examinee groups on which reliability estimates were based include students like those with whom you intend to use the test? | ☆ | ☆ | ☆ | ☆ | ☆ | ☆ | ☆ | ☆ |

**KEY**

☆ Important considerations which should be addressed.
√ Can be important under certain conditions.
† In cases where timed tests are used, technical assistance may be needed to interpret internal consistency estimates.

5. The test manual should indicate to what extent test scores are stable: that is, how nearly constant scores are likely to be if the test is administered after time has elapsed. (The time interval to be considered depends on the nature of the test and the recommended interpretation of test scores.)


Selected Test Publisher Responsibilities Regarding Test Reliability
• report evidence of reliability
• report reliability for all scores
• describe procedures and sample
• report information of alternate test forms
• report consistency over time

## Test Validity

*I had always thought of test validity as a fairly straightforward issue—that a test was either valid or it wasn't. Julie's comments gave me a new perspective on validity. Here's what she said:*

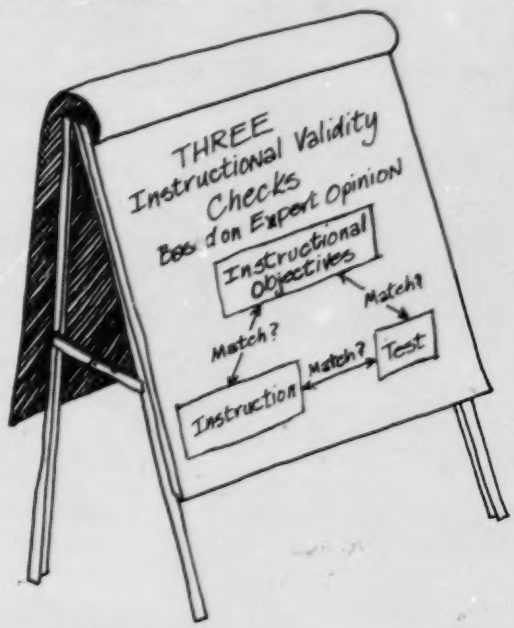
"Test validity depends on two factors: characteristics of the test itself and the manner in which the test is to be used. In other words, a test can be judged valid or invalid only in terms of its intended purpose: for example, diagnosis, selection, or program evaluation."

TYPES OF VALIDITY "Like reliability, validity has many subtle meanings. We are going to ignore most of those meanings and deal only with the two most useful in evaluating published tests: instructional validity and statistical validity.

"The *instructional validity* of a test is determined by relating three factors: (1) the objectives or intended outcomes of instruction, (2) the actual content of instruction, and (3) the objectives and content covered by the test. Very simply, if we as teachers decide what skills and content we want to teach, but the instruction we present doesn't match our intended outcomes, that instruction is invalid for achieving those outcomes. Similarly, if a test is intended to measure the extent to which students master the intended outcomes of instruction, and the test doesn't cover the course outcomes, it is invalid for its intended purpose.

"Test publishers typically present evidence in the test manual and/or user's guide that the skills and content covered in their tests are important and that the test items cover the intended content and skills. This evidence (sometimes referred to as *content validity*) is often simply the judgments of qualified subject matter and testing experts that the test is valid as a measure of intended outcomes.

**"Now, here is the important point:** ***Just because the publisher can prove that the test measures what he or she intended it to measure does not mean that it measures what you want to measure.*** There has to be a link between (1) your predetermined instructional objectives and actual instructional programs and (2) the content and skills measured by the test. Otherwise the test is not valid as a measure of *your* instructional objectives. No test should be administered until that link has been established.

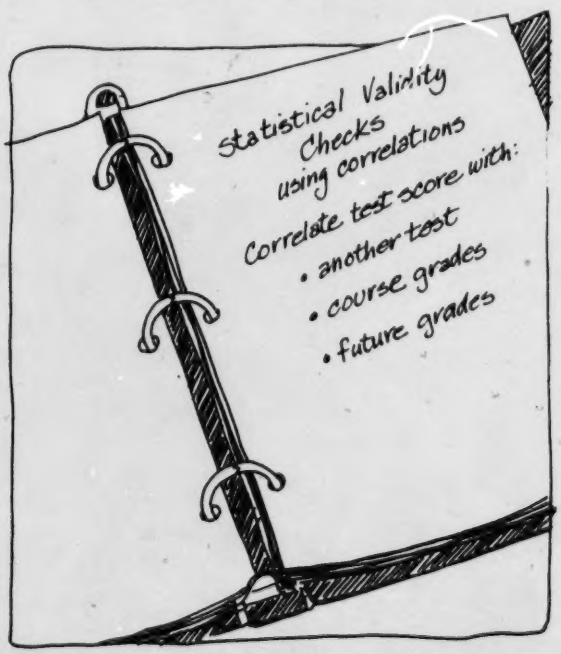



*Having given us some background information on instructional validity, Julie proceeded with a discussion of statistical validity.*

"A second way to determine test validity is to show that the score attained on the test by a given student suggests the same level of proficiency as do other indicators of performance for that student. Those 'other indicators' could be other established tests covering the same content. Or they could be performance on the job or in other real life settings—or classroom performance. In other words, if a student's test score

is consistent with or predictive of his or her performance in other settings, that's a good sign the test is a valid measure of what it's meant to measure. Let's consider some examples. If a student does well on an editing test and also performs well as editor of the school paper, there's evidence the test validly measures editing skill. If a student does well on a basic chemistry test but fails Basic Chemistry, the test's validity is called into question. Of course, many factors can influence grades or performance on the job. But the relationship between that performance and test performance is an important indicator of validity.

"Keep these three things in mind: First, no test is valid or invalid in and of itself. Rather, a test is valid or invalid for a given purpose. Second, comparing a student's test performance with some other measure of proficiency—another test, or grade—provides a significant indicator of validity. And third—here is where statistics come in—in each case, the measure of validity is the statistical relationship between the test in question and the other performance indicator. That's it—nothing complicated or hard to understand. Test publishers will refer to correlations between their tests and other performance indicators to defend the validity of their tests. We refer to these correlations as statistical reflections of validity. Common labels used by publishers are *construct*, *concurrent*, and *predictive* validity."




EVALUATING TEST VALIDITY "So how do you evaluate test validity? First, expect to encounter two general types of information about validity: (1) expert opinions regarding the appropriateness of test content and (2) correlation coefficients. Be sure the opinions offered are those of qualified, knowledgeable experts. Be a critical consumer and demand appropriate, understandable evidence. Make certain the test review process includes examination for and elimination of overtly sexist and racially biased items. And, *most important*, remember that you are responsible for determining that the content covered by the test matches the content you wish to measure.

"Interpreting correlation coefficients can be tricky. If you don't understand them, don't hesitate to get technical help. Generally, correlations range from zero to 1.00. Zero reflects no relationship at all, or complete inconsistency, between two test scores. At the other end of the continuum, 1.00 reflects perfect consistency between two scores. As a general guideline publishers report coefficients ranging upwards from .70 and clustering around .90. Though correlations below .70 may be acceptable, you should get some professional advice in interpreting correlations at or below that level. As a rule, the closer to 1.00, the better."

## Validity and Test Purpose

"There's more to evaluating test validity than just reading and evaluating data presented by the test publisher. Remember that the validity of a test depends on how it is to be used. Therefore, both general types of validity—instructional and statistical—must be evaluated in terms of your particular testing purpose. Let's consider how we might evaluate the validity of a test across our eight testing purposes.

"Remember that, in *diagnosis*, the purpose for testing is to pinpoint students' strengths and weaknesses so instruction can be planned. In this case teachers can best judge the instructional validity of a test, because they know the intended outcomes and the actual scope of instruction. Statistical validity could be established by showing that test performance was consistent with performance on another test covering the same content.

"The content validity of a test used for *placement* can best be judged by teachers familiar with the courses into which students will be placed. Statistical validity could be determined by establishing the relationship between test scores and grades received by students who had completed the course.

"*Guidance* tests often cover relatively broad content rather than specific objectives. Therefore, the subject matter experts in this context are those familiar with the content offered nationally. Statistical validity is often equated with the test's capacity to predict future success—grades in a vocational training program, for instance.

"Virtually the same is true of tests used in *selection*. Subject matter experts and teachers must judge the instructional validity. The index of statistical validity is test score correlation with subsequent performance—that is, grades—in the selected program.

"The teachers who planned instruction can best judge a test's instructional validity for *certification*. However, the prior decision of what objectives to test will be influenced by the opinions, and in some cases set policies, of school board members, legislators, local taxpayers, parents and future employers of the school's graduates. Another test covering the same material can serve as the basis for judging statistical validity.

"In the case of tests used for survey assessment educators, taxpayers and policy makers all should have a voice in determining what is tested. Further, as the size of the student group assessed expands—from building to district to state level—the size of the group judging instructional validity should also expand. Statistical validity can be determined via another test.

"For both *formative and summative evaluation*, those program personnel—administrators and teachers—most familiar with program offerings should judge test content appropriateness. In the case of summative evaluation, if the person most familiar with program content is not on the staff, that person's opinion should be sought anyway. Here again, a test or other performance indicator can serve to define statistical validity.

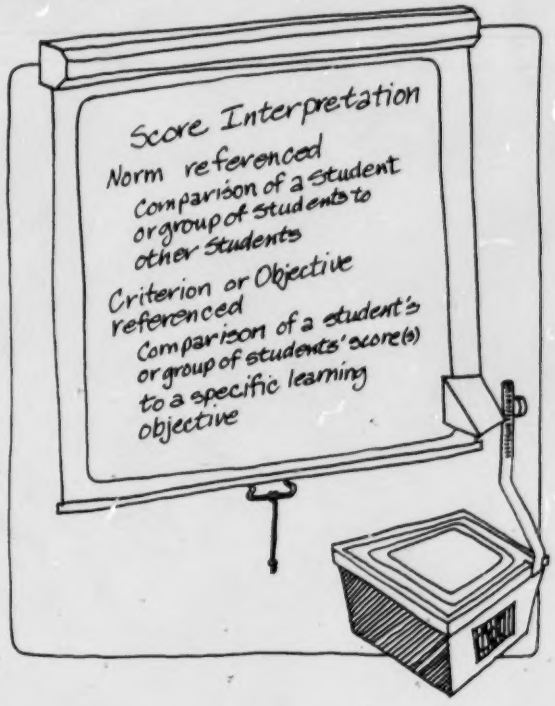
## Validity and Score Interpretation

*Julie concluded the validity discussion with a few brief comments on the interpretation of test scores, focusing on the difference between norm referenced and objective referenced tests.*

"First, in some cases, it is useful to compare an examinee's test performance with that of other students. Such comparisons are useful in selection, placement, and guidance. Ranking candidates might be one way of determining who will receive a scholarship, for instance. Comparing examinees is also useful in survey assessment; the adequacy of a target population's performance is often judged in light of how others performed on the same measure. In these contexts, the performance of a *norm* group is used as a basis for judging others' performance; hence the term *norm referenced interpretation*. Examples of norm referenced test scores include (1) percentiles, a number reflecting the percentage of the reference group whose performance a given examinee equaled or exceeded; and (2) grade equivalents, which suggest whether an examinee was performing below, at or above the average for his or her grade level. There are numerous other examples of norm referenced scores used by test publishers, and all have in common the comparison of performance among examinees.

"Three considerations are critical in selecting a norm referenced test for local use: (1) similarity between the norm group(s) and proposed examinees; (2) size of the norm group (it should be large enough to be truly representative); and (3) the levels of normative data available (building- or district-level norms should be available, if needed).

"In other contexts, a different standard of comparison is more relevant. For example, if an educational program is designed to help students attain a certain level of proficiency on given instructional objectives, it is irrelevant how students perform relative to one another. What matters is how they perform relative to each of the established instructional objectives. This way of judging performance is termed *objective referenced or criterion referenced interpretation*. For diagnosis, formative evaluation, and certification, it is common to establish very specific learning outcomes. Test data address the question: Have examinees mastered the required skills? Test items are frequently tied to specific instructional objectives, making it easier to determine an examinee's proficiency on each objective.


Score Interpretation
Norm referenced
Comparison of a student or group of students to other students
Criterion or Objective referenced
Comparison of a student's or group of students' score(s) to a specific learning objective

"Successful selection of a criterion or objective referenced test depends on careful consideration of instructional validity. Test objectives must match intended instructional outcomes. Equally important, the criterion level for judging mastery must be appropriate. For example, if 10 items are used to assess one objective, how many must the examinee answer correctly to be judged proficient? Five? Eight? All ten? Generally, a cutoff score is established through expert review of items and objectives. In other words, the decision is subjective. The informed test consumer will therefore scrutinize the credentials of those who

set these criterion levels. And here, as in other cases, you should seek professional assistance as needed."

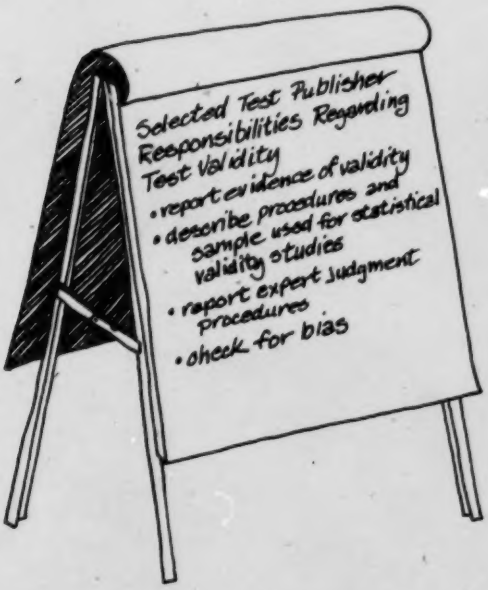
*In conclusion, Julie provided us with another summary table covering validity (Table 3) and reminded the group that test publishers are obligated to provide users with certain information about validity. Here are a few standards for judging the quality of that evidence:*

1. A manual or research report should present evidence of validity for each context in which use of the test is recommended. If validity for some suggested interpretation has not been investigated, that fact should be made clear. Test *users* are responsible for marshalling the evidence in support of their validity claims. The use of test scores in decision rules should be supported by evidence.

2. The sample employed in a validity study and the conditions under which testing is done should be consistent with recommended test use and should be described sufficiently for the reader to judge their pertinence to his or her situation.

3. The manual should describe the relevant professional experience and qualification of any experts asked to review test items and procedures, as well as the directions under which they made their judgments.

4. The publisher should provide evidence that test items have been properly and thoroughly reviewed for bias.


Selected Test Publisher Responsibilities Regarding Test Validity
• report evidence of validity
• describe procedures and sample used for statistical validity studies
• report expert judgment procedures
• check for bias

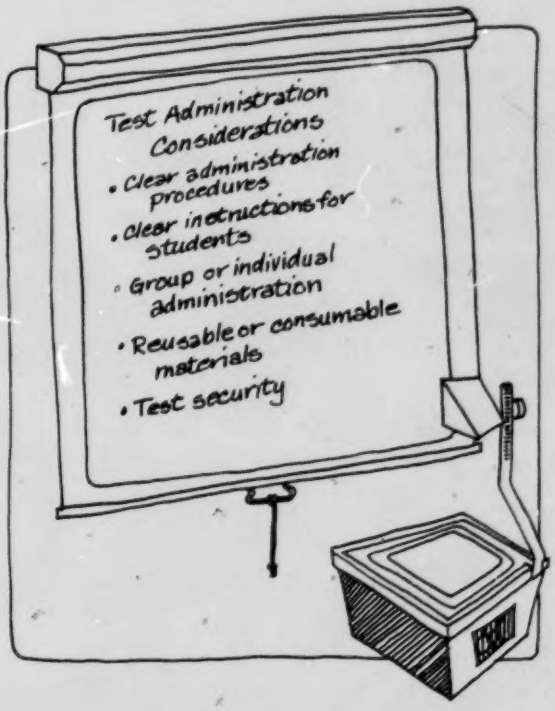
## The Logistics of Testing

*The second day of the workshop was devoted to the logistics of testing. Julie presented some very practical ideas:*

"Good tests yield reliable and valid scores *efficiently*, with a minimum investment of your time and money. As with reliability and validity, some aspects of testing efficiency and cost remain constant, while others vary as a function of testing purpose."

TEST ADMINISTRATION "Under any circumstances, the test publisher is obligated to provide (1) a complete description of the test administration characteristics and (2) a clear and complete set of verbatim instructions for administering the test. Administration procedures partially determine how useful a test will be in various contexts. For example, tests may require group or individual administration or they may rely on reusable or consumable test materials. These factors should be covered in test administration materials provided by publishers.

"Administration instructions should cover procedures for preparing to test, and instructions for test scoring, as well as verbatim instructions to examinees. Without information in each of these areas, testing will be difficult—if not impossible—regardless of the assessment context.


Test Administration Considerations
• Clear administration procedures
• Clear instructions for students
• Group or individual administration
• Reusable or consumable materials
• Test security

"Factors that vary by context are (1) the appropriateness of various modes of test administration and (2) the relative importance of test security. Let's con-

Table 3: **Key Elements in Evaluating Test Score**

## VALIDITY

**as a Function of Test Purpose**

| Important Questions to be Addressed | INSTRUCTIONAL MANAGEMENT | | | SCREENING | | PROGRAMMATIC DECISIONS | | |
|---|---|---|---|---|---|---|---|---|
| | Diagnosis | Placement | Guidance/ Counseling | Selection | Certification | Survey Assessment | Formative Evaluation | Summative Evaluation |
| **General Question:** Can we depend on the scores to serve their intended purpose? | ☆ | ☆ | ☆ | ☆ | ☆ | ☆ | ☆ | ☆ |
| The published test manual and your test planning should address the following **more specific** questions: | | | | | | | | |
| 1. Does the test publisher present evidence of the appropriateness of the content covered (e.g. qualified reviewers' opinions presented appropriately)? | ✓ Appropriateness for local priorities more crucial | ✓ Appropriateness for local course content more crucial | ☆ | ☆ | ✓ Appropriateness for local priorities more crucial | ☆ | ✓ Appropriateness for local priorities more crucial | ✓ |
| 2. Does the publisher verify the link of test items to the intended focus of the test? | ☆ | ☆ | ☆ | ☆ | ☆ | ☆ | ☆ | ☆ |
| 3. As a test user, have you verified the appropriateness of test content coverage to your own local curricular priorities? | ☆ | ☆ | ☆ | ☆ | ☆ | ☆ | ☆ | ☆ |
| 4. Does the test publisher present evidence that the test was reviewed by disinterested, external test reviewers to ensure the educational and social (unbiased) appropriateness of the test items? | ☆ | ☆ | ☆ | ☆ | ☆ | ☆ | ☆ | ☆ |
| 5. Will the content specifications of the test and the actual test items appear appropriate and useful to those who are to use the data to make educational decisions? | ☆ | ☆ | ☆ | ☆ | ☆ | ☆ | ☆ | ☆ |
| 6. Is the validation group similar to the test group? | ☆ | ☆ | ☆ | ☆ | ☆ | ☆ | ☆ | ☆ |
| 7. Are scores on other tests covering the same material correlated with scores on the test in question? | ☆ | ☆ | ☆ | ☆ | ☆ | ☆ | ☆ | ☆ |
| 8. Is the predictive power of the test established? | ✓ Prediction may not be an issue | ☆ | ☆ | ☆ | ✓ | ✓ Prediction may not be an issue | ✓ | ✓ |
| **Test Score Interpretation** | | | | | | | | |
| 1. Are scores sufficiently precise to help with the decision to be made? | ☆ | ☆ | ☆ | ☆ | ☆ | ☆ | ☆ | ☆ |
| 2. Are norm groups described and appropriate? | ✓ Generally criterion referenced | ☆ | ☆ | ☆ | ✓ Generally criterion referenced | ☆ | ✓ Generally criterion referenced | ☆ |
| 3. Are mastery scores established and appropriate? | ☆ | ✓ Local cutoff scores | ✓ Generally norm referenced | ✓ | ☆ | ☆ | ☆ | ☆ |

**KEY**

☆ Important considerations which should be addressed.

✓ Can be important under certain conditions.

sider test administration first.

"Tests requiring individual administration (one examiner to one examinee) are less usable in large-scale testing than in diagnosis or individual guidance. In addition, tests relying on reusable test materials may be most useful in large-scale testing or diagnosis, where tests are likely to be administered more than once. Reusable materials create unnecessary expense for one-time administration.

"The amount of test security required to ensure fairness also varies by context. Test security is almost totally the province of the user. Tight test security is most important where the likelihood of negative consequences to individuals is high—for instance, admission into a desired program, or certification of minimum competence for high school graduation. In formative program evaluation, guidance, and diagnosis, security is less crucial.

"In considering the issues of reusable test materials and test security, it is sometimes useful to consider the option of returning the test to the students and their parents for review. When the test is not to be readministered, it may be helpful for the student to study and learn from items answered incorrectly. Similarly, you can use such tests to communicate to parents the kind(s) of material the student has or has not mastered.

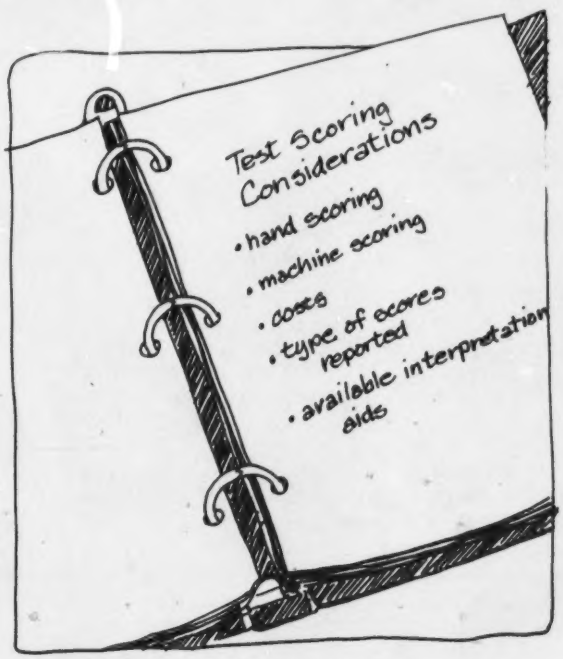
"When tight security is a must, the best tests are those which (1) have not been circulated or used at other times with the target population, (2) are available in equivalent forms for multiple test administration, and (3) do not require costly test booklets (since security may not allow booklet reuse). On the other hand, the economy of reusable materials is a great advantage when security is not an issue."

TEST SCORING AND REPORTING "Careful test selection includes a careful review of the scoring services offered by the publisher or other accessible agencies. Regardless of context, the publisher is obligated to provide a clear and complete set of scoring instructions, including procedures for converting scores to any interpretive—for example, norm referenced—scores. In addition, the publisher should spell out in detail the alternative scoring and reporting services available to the user and the associated costs (in time and money). Specifications should cover both hand and machine scoring and any options available in score reporting.

"Machine scoring is generally preferable to hand scoring any time there are large numbers of tests to be scored. Because machine scoring is much faster—and therefore more economical—than hand scoring, it is highly appropriate for most large-scale test administrations, regardless of context. However, hand scoring may be a viable alternative in small-scale assessment as part of formative evaluation or diagnosis. Hand scoring can be done immediately, producing timely feedback. And sometimes, patterns of incorrect responses can be spotted by teachers, thus providing a kind of diagnostic information. In addition, hand scoring may be necessary with some types of tests—writing samples, for instance—that cannot be machine scored.

"Context largely determines the best approach to score reporting. Generally, evaluation and diagnosis require a finer breakdown of results than guidance and selection. The scoring service selected should be capable of producing precisely the data needed. The costs and problems associated with re-analyzing results after the fact can be extensive.

"Many test publishers offer printed—and sometimes audiovisual—test score interpretation aids. You may find such aids useful with the more complicated scores used in norm referenced interpretation—particularly in selection, guidance, and placement. Interpretive aids may be helpful in these cases. Aids are less necessary in criterion referenced contexts involving straightforward mastery decisions."




COSTS OF TESTING "Regardless of context, assessment costs generally comprise costs of four components: test booklets, answer sheets, scoring services and the time required to conduct the assessment—including training and preparing test administrators, administering the test, scoring and reporting results, and interpreting and using the results. No testing program should be initiated until all costs have been accurately calculated and weighed against the value of the results to be sure that the assessment is a wise investment."

*Julie summarized testing logistics for us (Table 4) and offered these few guidelines for dealing with testing logistics:*

1. Establish communication with test publisher sales representatives in your area and work closely with them throughout test selection. They are a valuable source of relevant information.

2. Always comparison-shop. Consider alternative test publishers and alternative vendors for such services as test scoring. Testing is a very competitive market; take advantage of that fact.

3. Always spot-check machine scoring services by hand scoring *a few* answer sheets and comparing the results with those of the scoring service. Though scoring services are highly reliable, they do make mistakes and a few moments spent verifying scores can help ensure that those rare mistakes do not jeopardize your assessment.

4. Spot-check the score summaries provided by computerized scoring services. With very large assessments, this may not be feasible. But at the local level, a few moments spent in checking can prevent problems later.

5. Seek the professional advice of a competent measurement expert whenever you have a question or problem you cannot handle. Questions left unaddressed can impair the effectiveness of an otherwise carefully planned and conducted assessment.

## REFERENCES FOR TEST SELECTION

*Having gained some background on the interrelationships between testing purposes and characteristics, we were ready for specific information on basic skills and life skills tests. Julie provided a list of currently available tests in both areas, and suggested some reference books for further information on the tests. These references are included in the appendices:*

- *Appendix A—Currently available basic skills and life skills tests*
- *Appendix B—Publishers' names, addresses, telephone numbers*
- *Appendix C—Reference materials describing and reviewing each test*

*Suppose you have a specific reason for testing—let's say evaluating a reading program. After reviewing the list of available tests, you may be able to limit your preliminary selection to five or six that seem most appropriate. Write to the publishers. Most sell "specimen sets" which contain samples of the tests and other information needed to decide which test is best for a particular testing purpose. Also, don't hesitate to obtain additional technical assistance in test selection.*

## Table 4: Key Elements of TEST LOGISTICS as a Function of Test Purpose

| Important Questions to be Addressed | INSTRUCTIONAL MANAGEMENT | | | SCREENING | | PROGRAMMATIC DECISIONS | | |
|---|---|---|---|---|---|---|---|---|
| | Diagnosis | Placement | Guidance/ Counseling | Selection | Certification | Survey Assessment | Formative Evaluation | Summative Evaluation |
| **General Question:** Are the procedures and costs associated with test implementation appropriate, given the purpose for testing? | ☆ | ☆ | ☆ | ☆ | ☆ | ☆ | ☆ | ☆ |
| The published test manual and your test planning should address the following **more specific** questions: | | | | | | | | |
| **Test Administration** | | | | | | | | |
| 1. Are test administration characteristics clearly defined (e.g., timed vs. untimed; individual vs. group, etc.)? | ☆ | ☆ | ☆ | ☆ | ☆ | ☆ | ☆ | ☆ |
| 2. Are complete verbatim instructions for test administration provided? | ☆ | ☆ | ☆ | ☆ | ☆ | ☆ | ☆ | ☆ |
| 3. Is the test designed for group administration? | ✓ Individual may be acceptable | | ✓ | ☆ | ☆ | ☆ | ☆ | ☆ |
| 4. Are test booklets designed to be reusable? | ☆ | ☆ | ✓ Not important if one-time administration | ✓ | ✓ | ☆ | ☆ | ☆ |
| 5. Is test security an important issue? | ✓ | ✓ | ✓ | ☆ | ☆ | ✓ | ✓ | ✓ |
| **Test Scoring** | | | | | | | | |
| 1. Are clear and complete instructions for scoring included (i.e., hand scoring, machine scoring, or both)? | ☆ | ☆ | ☆ | ☆ | ☆ | ☆ | ☆ | ☆ |
| 2. Are score interpretation procedures provided? | ☆ | ☆ | ☆ | ☆ | ☆ | ☆ | ☆ | ☆ |
| 3. Are machine-scoring alternatives available, or must the user hand-score the tests? | ✓ With small examinee group, hand-scoring may be viable | | ✓ | ☆ | ☆ | ☆ | ✓ With small examinee group, hand scoring may be viable | ☆ |
| 4. Are scores reported in a manner that will make them directly usable in the decision making process? | ☆ | ☆ | ☆ | ☆ | ☆ | ☆ | ☆ | ☆ |
| 5. Are interpretive aids available? | ✓ Less necessary with criterion referenced tests | ☆ | ☆ | ☆ | ✓ Less necessary with criterion referenced tests | ☆ | ✓ Less necessary with criterion referenced tests | ✓ |

**KEY**

☆ Important considerations which should be addressed.

✓ Can be important under certain conditions.

**CONTINUED ON FOLLOWING PAGE**

Table 4: Key Elements of
## TEST LOGISTICS
as a Function of Test Purpose

| Important Questions to be Addressed | INSTRUCTIONAL MANAGEMENT | | | SCREENING | | PROGRAMMATIC DECISIONS | | |
|---|---|---|---|---|---|---|---|---|
| | Diagnosis | Placement | Guidance/ Counseling | Selection | Certification | Survey Assessment | Formative Evaluation | Summative Evaluation |
| **Costs of Testing** | | | | | | | | |
| 1. Have all cost factors associated with testing been estimated, including: | ✭ | ✭ | ✭ | ✭ | ✭ | ✭ | ✭ | ✭ |
| Cost of test booklets | | | | | | | | |
| Cost of answer sheets | | | | | | | | |
| Cost of scoring | | | | | | | | |
| Cost, in time, of | | | | | | | | |
| (1) training users | | | | | | | | |
| (2) administering tests | | | | | | | | |
| (3) obtaining scores | | | | | | | | |
| (4) interpreting and using scores | | | | | | | | |

**KEY**
✭ Important considerations which should be addressed.
✓ Can be important under certain conditions.

## SUMMARY AND FOLLOWUP

*In summary, our awareness workshop covered the relationship between testing purposes and characteristics, and provided references on published basic and life skills tests. To review, the potentials and merits of any test can openly be judged in light of the proposed purpose for testing. Testing purposes relate to eight educational decision-making contexts:*

- *Instructional management decisions*
  1. *Diagnosis*
  2. *Placement*
  3. *Guidance and Counseling*
- *Student Screening Decisions*
  4. *Selection*
  5. *Certification*
- *Programmatic Decisions*
  6. *Survey Assessment*
  7. *Formative Evaluation*
  8. *Summative Evaluation*

*In light of your testing purpose, be sure to investigate these test characteristics:*

- *Test Score Consistency (reliability)*
- *Test Validity*
- *The Logistics and Costs of Testing*

*Remember, many helpful reference materials exist to guide you in test evaluation and selection.*

*Planning a testing program requires the involvement of people throughout the educational system and the community. Be sure your testing plan specifies every step from identifying testing purposes to interpreting and reporting results. Only then will you have a solid foundation for test selection.*

*Our first meeting with Julie provided a strong foundation for planning the district testing program. Luckily, our enthusiasm carried us through several months of hard work. During that time we accomplished a number of critical tasks.*

*Our first task was to outline the main testing purposes for the district. After much discussion among ourselves and with district and school staff, we established the following testing purposes and schedule.*

*District wide survey assessment was to be conducted in the spring in grades 3, 6, 9 and 11. The grade 6 survey assessment was to cover the same skills as the statewide survey assessment test given at grade 6. In addition, the grade 6 mathematics scores were to be used to place students in junior high math classes.*

*There were two district wide programs which would require evaluation. One was the federally funded reading program for disadvantaged students. We decided that summative evaluation tests would be needed for this program to determine if the program was doing the job. Tests which were used for the selection of students to participate in the program and for formative program evaluation were already in use. We decided to recommend that these tests be used along with a final test at the end of the program to evaluate program impact.*

*The other program, a junior high school social studies curriculum being pilot tested in one high school, also needed a summative evaluation test. For this project, we recommended that tests should be given to program participants at the end of grades 10 and 11.*

*We decided that diagnostic testing should remain the responsibility of individual teachers because of their close day-to-day contacts with students and instruction. Therefore, we made no recommendations regarding this purpose. We also recognized that the college entrance exams were useful guidance tools for students and parents, so we didn't recommend any additional guidance testing. However, we did recommend further exploration of the career planning assessment area by the guidance staff (e.g., interests, etc.).*

*Once we delineated testing purposes and schedules, we spent time checking our recommendations with district staff, administrators and the school board. On the whole, they agreed with us, so we moved on to selecting tests. Since we had tried to make our tests serve multiple purposes, we found that test selection was a series of compromises. For example, we tried to combine district and state survey assessment. Thus, we ended up selecting the statewide test items for our district survey assessment whenever possible. Most people were happy with the resulting tests, so that caused no problem. We also wanted to use the mathematics section of the sixth grade test for placement of students in junior high math courses. But math teachers felt that the assessment test gave them inadequate information for student placement. Rather than add another test, however, they developed a system to use test scores, grades, and referrals from sixth grade teachers in the placement process.*

*Once tests were selected, we thought we had finished our job and, with sighs of relief, we were about to disband the committee. However, the superintendent pointed out that with just a bit more work, we could bring our committee work to an even more useful conclusion. At that point we went to work on a series of meetings with Julie regarding the interpretation and effective reporting of test scores. But that's another story to be told at another time. Suffice it to say, we recommended not only what test to use and when to give the test, but also who should receive test results, what scores they should receive, and how the test results should be reported.*

*This year is the first year of the new testing program. So far most people are pleased with the results and I think they see the tests as really useful (for a change). Later in the year, we plan to convene a test committee reunion to evaluate the results of our work and recommend any necessary changes.*

# APPENDIX A

## Currently Available Basic Skills and Life Skills Tests

To assist test users in beginning the process of selecting published tests for use in their particular settings, a list of 102 published tests compiled from a number of sources is included here. These tests are intended to be broadly representative of the domain of published tests measuring basic and life skills. Although we attempted to include all tests in these categories published since 1970, some may have been inadvertently overlooked. Included in the list are 29 multi-subject achievement test batteries, 35 standalone tests of reading ability, 18 tests of mathematics ability, 15 life skills tests and five tests dealing only with language arts skills. *The inclusion of a test in the list does not imply endorsement.*

The lists are intended to give the user a starting place for test selection. Therefore, detailed information is not provided. Rather, the information provided includes: Test name and names of subtests (if appropriate), grade levels for which the test can be used, publication date, publisher and sources of additional information about the test.

To facilitate follow-up steps in the test selection process, a list is provided in Appendix B of the names, addresses and telephone numbers of the publishers whose tests are cited. Each of these reference documents is discussed in Appendix C:

**MMY** Buros, Oscar, K., Ed. *The Eighth Mental Measurements Yearbook*. 1978. The Gryphon Press, Highland Park, NJ.

**TIP** Buros, Oscar, K., Ed. *Tests in Print II*. 1974. The Gryphon Press, Highland Park, NJ.

**FLIT** Nafziger, Dean H.; Thompson, R. Brent; Hiscox, Michael D. and Owen, Thomas R. *Tests of Functional Adult Literacy: An Evaluation of Currently Available Instruments*. 1975. Northwest Regional Educational Laboratory, Portland, OR 97204.

**NCME** "New from Test Publishers," *NCME Measurement News*. Early Winter '77–Spring '79. National Council on Measurement in Education, 1200 17th Street N.W., Washington, DC 20036. (Periodical).

**NOT** "News on Tests," Jan. '79–Nov. '79 Test Collection, Educational Testing Service, Princeton, NJ 08541. (Periodical).

**TCB** *Test Collection Bulletin*. Jan. '77–Oct. '78. Test Collection, Educational Testing Service, Princeton, NJ 08541. (Periodical).

## MULTISUBJECT ACHIEVEMENT BATTERIES

| Tests and Subscores* | Grade Level(s) | Publication Date | Publisher** | Reference*** |
|---|---|---|---|---|
| Adult Basic Learning Examination (ABLE)<br>Reading<br>Arithmetic<br>Spelling<br>Vocabulary | Adult | 1974 | Psy. Corp. | MMY 2 |
| Adult Performance Level Survey (APLS)<br>Reading<br>Computation<br>Writing<br>Identifying Facts and Terms<br>Problem Solving | 9-Adult | 1976 | ACT | MMY 3 |
| Alaska Instructional Diagnostic System (AIDS)<br>Reading<br>Mathematics | 1-8 | 1977 | SRRC | TCB July 78 pg. 6 |
| American School Achievement Tests, Revised Edition (ASAT)<br>Reading<br>Arithmetic<br>Language<br>Spelling<br>Social Studies<br>Science | 1-9 | 1975 | BMC | MMY 4 |
| Basic Educational Skills Inventory (BESI)<br>Reading<br>Mathematics | Kindergarten-6 | 1973 | Winch | MMY 5 |
| Brigance Diagnostic Inventory of Basic Skills (IBS)<br>Reading<br>Mathematics<br>Language Arts<br>Readiness | Kindergarten-6 | 1976 | CAI | MMY 6 |
| California Achievement Tests Forms C & D (CAT)<br>Reading<br>Mathematics<br>Language<br>Spelling<br>Reference Skills | Kindergarten-12 | 1977 | CTB | TCB July 77 pg. 10 |
| California Assessment Program Survey of Basic Skills<br>Reading<br>Mathematics<br>Written Expression<br>Spelling | 12 | 1974 | CSDE | TCB July 78 |

CONTINUED ON FOLLOWING PAGE

*Note: Not all subtests are available at all grade levels
**Complete address and phone number in Appendix B
***Reference documents described in detail in Appendix C

## MULTISUBJECT ACHIEVEMENT BATTERIES

| Tests and Subscores | Grade Level(s) | Publication Date | Publisher | Reference |
|---|---|---|---|---|
| CIRCUS, Levels C & D<br>Reading<br>Mathematics<br>Writing Skills<br>Listening<br>Phonetic Analysis<br>Oral Reading<br>Say and Tell<br>Do You Know<br>Think It Through<br>Things I Like<br>Educational Environment Questionnaire | 1-3 | 1979 | AW | NOT Sept 79 |
| Comprehensive Tests of Basic Skills Expanded Edition Forms S&T (CTBS)<br>Reading<br>Mathematics<br>Language Arts<br>Reference Skills<br>Science<br>Social Studies | Kindergarten-12 | 1976 | CTBS | MMY 12 |
| Criterion Test of Basic Skills<br>Reading<br>Arithmetic | Kindergarten-8 | 1976 | ATP | MMY 14 |
| Diagnostic Skills Battery<br>Reading<br>Mathematics<br>Language Arts | 1-8 | 1976 | STS | TCB Jan 77 pg. 5 |
| Iowa Tests of Basic Skills Multi-level Edition Forms 7&8<br>Reading Comprehension<br>Mathematics Skills<br>Language Skills<br>Work-Study Skills<br>Vocabulary | 3-9 | 1978 | HM | NOT July 79 |
| Iowa Tests of Educational Development: SRA Assessment Survey<br>Reading<br>Mathematics<br>Language Arts<br>Social Studies<br>Science | 9-12 | 1974 | SRA | MMY 20 |
| Metropolitan Achievement Tests (METRO '78)<br>Reading Comprehension<br>Mathematics<br>Language<br>Social Studies<br>Science | Kindergarten-12 | 1978 | Psy. Corp. | Fall 78 NCME |

CONTINUED ON FOLLOWING PAGE

## MULTISUBJECT ACHIEVEMENT BATTERIES

| Tests and Subscores | Grade Level(s) | Publication Date | Publisher | Reference |
|---|---|---|---|---|
| National Educational Development Tests<br>Mathematics Usage<br>English Usage<br>Social Studies Reading<br>Natural Sciences Reading<br>Word Usage | 7-10 | 1974 | SRA | MMY 23 |
| Primary Survey Tests<br>Reading<br>Mathematics<br>Language<br>Spelling | 2-3 | 1973 | SF | TIP 27 |
| Scholastic Testing Service<br>Educational Development Series<br>Scholastic Tests<br>Reading<br>Mathematics<br>English<br>Social Studies<br>Science<br>Solving Everyday Problems<br>USA in the World<br>Nonverbal Ability<br>Verbal Ability<br>School Interests<br>School Plans<br>Career Plans | 2-12 | 1976 | STS | MMY 20 |
| Science Research Associates<br>Achievement Series (ACH) Forms 1&2<br>Reading<br>Mathematics<br>Language Arts<br>Social Studies<br>Science<br>Reference Materials<br>Applied Skills | Kindergarten-12 | 1978 | SRA | NCME Fall 78 |
| Science Research Associates<br>High School Placement<br>Reading<br>Arithmetic/or Modern Math<br>Language Arts<br>Social Studies<br>Science | 9 | 1973 | SRA | TIP 31 |
| Science Research Associates<br>Norm Referenced/Criterion Referenced Testing Program<br>Reading<br>Mathematics | 4-10 | 1977 | SRA | TCB July 77 |

CONTINUED ON FOLLOWING PAGE

## MULTISUBJECT ACHIEVEMENT BATTERIES

| Tests and Subscores | Grade Level(s) | Publication Date | Publisher | Reference |
|---|---|---|---|---|
| Sequential Tests of Educational Progress: Series III Levels E-J† | 1-12 | 1979 | AW | NOT Sept 79 |
| Reading | | | | |
| Mathematics Computation | | | | |
| Mathematics Basic Concepts | | | | |
| Writing Skills | | | | |
| Social Studies | | | | |
| Science | | | | |
| Study Skills/Listening | | | | |
| Goal Orientation Index | | | | |
| SOI Learning Abilities Test | 1-11 | 1975 | SOI | NOT April 79 |
| Reading | | | | |
| Arithmetic | | | | |
| Stanford Achievement Test 1973 Edition (SAT) | 1-9 | 1973 | Psy Corp. | MMY 29 |
| Reading | | | | |
| Mathematics | | | | |
| Social Studies | | | | |
| Science | | | | |
| Listening Comprehension | | | | |
| Spelling | | | | |
| Stanford Test of Academic Skills (TASK) | 8-Adult | 1975 | Psy Corp. | MMY 31 |
| Reading | | | | |
| Mathematics | | | | |
| English | | | | |
| Tests of Adult Basic Education (TABE) | Adult | 1976 | CTB | MMY 33 |
| Reading | | | | |
| Mathematics | | | | |
| Language Arts | | | | |
| Tests of Achievement and Proficiency, Form T (TAP) | 9-12 | 1978 | RS | NOT July 79 |
| Reading Comprehension | | | | |
| Mathematics | | | | |
| Written Expression | | | | |
| Social Studies | | | | |
| Science | | | | |
| Using Sources of Information | | | | |
| Applied Proficiency Skills | | | | |
| United States Employment Service Basic Occupational Literacy Test (USES BOLT) | Adult | 1974 | USDL | MMY 489 |
| Reading Vocabulary | | | | |
| Reading Comprehension | | | | |
| Arithmetic Computation | | | | |
| Arithmetic Reasoning | | | | |
| Wide Range Achievement Test Revised Edition (WRAT) | Kindergarten-Adult | 1978 | Jastak | NOT Sept 79 |
| Reading | | | | |
| Arithmetic | | | | |
| Spelling | | | | |

†Step III is integrated with CIRCUS C&D for the beginning of grade 1 through the beginning of grade 3.

## READING TESTS

| Test | Grade Level(s) | Publication Date | Publisher | Reference |
|---|---|---|---|---|
| Analysis of Skills: Reading | 1-8 | 1976 | STS | MMY 748 |
| Analytical Reading Inventory | 2-9 | 1977 | Merrill | TCB Jan 78 Pg. 2 |
| Clarke Reading Self-Assessment Inventory | 11-Adult | 1978 | ATP | NOT Feb 79 |
| Criterion Referenced: Reading Tactics | 7-12 | 1977 | SF | TCB July 77 Pg. 3 |
| Diagnostic Reading Scales | 1-12 | 1975 | CTB | MMY 753 |
| Diagnostic Screening Test: Reading | 1-12 | 1979 | SC | NOT Oct 79 |
| Diagnosis: An Instructional Aid: Reading | 1-6 | 1974 | SRA | MMY 752 |
| Diagnostic Reading Test: Pupil Progress Series | 1-8 | 1970 | STS | TIP 1625 |
| Fountain Valley Teacher Support System System in Reading (FVTSS-R) | 1-6 | 1975 | Zweig | MMY 725 |
| Fountain Valley Teacher Support System in Secondary Reading | 7-12 | 1976 | Zweig | MMY 758 |
| Gates-McGinitie Reading Tests, Second Edition | 1-12 | 1978 | RS | NOT Nov 79 |
| Gates-McGinitie Reading Tests, First Edition | 1-12 | 1970 | RS | TIP 1553 |
| Individual Pupil Monitoring System-Reading (IPMS-Reading) | 1-6 | 1974 | RS | MMY 763 |
| Individualized Criterion Referenced Testing: Reading (ICRTR) | K-8 | 1976 | EDC | MMY 764 |
| Mastery: An Evaluation Tool: Reading (SOBAR) | K-9 | 1976 | SRA | MMY 766 |
| McCarthy Individualized Diagnostic Reading Inventory-Revised | 2-Adult | — | EPS | TCB Jul 77 Pg. 4 |
| McGrath Diagnostic Reading Test | 1-13 | 1976 | McGrath | MMY 756 |
| The McGuire-Bumpus Diagnostic Comprehension Test (MBDCT) | 2-6 | 1972 | Croft | TIP 1635 |
| Nelson-Denny Reading Test Forms C&D (NDRT) | 9-12 | 1976 | RS | MMY 735 |
| Objectives-Referenced Bank of Items and Tests: Reading and Communication Skills (ORBIT: RCS) | K-Adult | 1975 | CTB | MMY 767 |
| Nelson Reading Skills | 3-9 | 1977 | RS | TCB Jul 77 Pg. 4 |
| Power Reading Survey Test | 1-12 | 1975 | BFA | MMY 768 |
| Prescriptive Reading Inventory | K-6 | 1977 | CTB | MMY 769 |
| Ransom Program Reading Tests | K-6 | 1975 | AW | MMY 770 |
| Reading Inventory Probe I | 1-2 | 1973 | ATC | TIP 1725 |
| Reading: IOX Objectives-Based Tests | K-6 | 1976 | IOX | MMY 771 |
| Reading Skills Diagnostic Test | 2-8 | 1971 | BP | MMY 772 |
| Reading Skills Competency Tests | K-7 | 1979 | CARE | NOT Oct 79 |
| Reading Unlimited Skills Assessment Tests | 1-6 | 1976 | SF | TCB Oct 78 Pg. 2 |
| SRA Reading Index | 9-12 & Adult | 1974 | SRA | MMY 813 |
| Skills Monitoring System: Reading (SMSR) | 3-5 | 1975 | Psy. Corp. | MMY 776 |

| | | | | |
|---|---|---|---|---|
| Stanford Diagnostic Reading Test (SDRT) | 1-13 | 1976 | Psy. Corp. | MMY 777 |
| Wisconsin Design for Reading Skill Development | K-6 | 1972 | NCS | MMY 778 |
| Woodcock Reading Mastery Tests (WRMT) | K-12 | 1973 | AGS | MMY 779 |

## MATHEMATICS TESTS

| Test | Grade Level(s) | Publication Date | Publisher | Reference |
|---|---|---|---|---|
| Analysis of Skills: Mathematics (ASK: Mathematics) | 1-8 | 1976 | STS | MMY 251 |
| Assessment of Skills in Computation (ASC) | 7-9 | 1978 | CTB | NOT Oct 79 |
| Basic Arithmetic Skill Evaluation | 1-9 | 1974 | IILC | MMY 303 |
| Diagnosis: An Instructional Aid: Mathematics | 1-6 | 1974 | SRA | MMY 263 |
| Diagnostic Mathematics Inventory (DMI) (Revision of the PMI) | 1-8 | 1975 | CTB | MMY 264 |
| Diagnostic Screening Test: Math (DSTM) | 1-11 | 1979 | SC | NOT Nov 79 |
| ERB Modern Arithmetic Test | 5-6 | 1971 | ERB | TIP 718 |
| Fountain Valley Teacher Support System in Mathematics (FVTSS-M) | K-8 | 1974 | Zweig | MMY 270 |
| Individualized Criterion Referenced Testing: Math (ICRTM) | 1-8 | 1977 | EDC | MMY 275-6 |
| Individual Pupil Monitoring System-Mathematics (IPMS) | 1-8 | 1973 | RS | MMY 274 |
| Keymath Diagnostic Arithmetic Test | K-6 | 1976 | AGS | MMY 305 |
| Mastery: An Evaluation Tool: Mathematics | K-9 | 1976 | SRA | MMY 278 |
| Mathematics: IOX Objectives-Based Tests | K-9 | 1976 | IOX | MMY 279 |
| Minimum Essentials for Modern Math | 6-8 | 1971 | Hayes | TIP 638 |
| Objectives-Referenced Bank of Items and Tests: Mathematics (ORBIT:M) | K-Adult | 1975 | CTB | MMY 287 |
| Stanford Diagnostic Mathematics Test (SDMT) | 1-Adult | 1976 | Psy. Corp. | MMY 292 |
| Steenburgen Quick Math Screening Test | 1-6 | 1978 | ATP | NOT Feb 79 |
| Tests of Achievement in Basic Skills: Mathematics (TABS:M) | K-12 | 1976 | EdITS | MMY 293 |

## LIFE SKILLS TESTS

| Test | Grade Level(s) | Publication Date | Publisher | Reference |
|---|---|---|---|---|
| Adult Performance Level Functional Literacy Test | 9-Adult | 1978 | ACT | FLIT Pg. 42 |
| Assessment of Skills in Computation (ASC) | 7-9 | 1978 | CTB | NOT Oct 79 |
| Everyday Skills Tests (EDST) | 6-12 | 1975 | CTB | MMY 18 |
| IOX Basic Skills Test | 9-12 | 1978 | IOX | NCME Spring 1979 |
| NM Consumer Mathematics Test | 9-12 | 1973 | NMS | MMY 312 |
| Reading/Everyday Activities in Life (R/EAL) | 9-Adult | 1972 | CAL-P | FLIT Pg. 46 |
| SRA Coping Skills: A Survey plus Activities | 7-8 & Adult | 1979 | SRA | NCME Special Edition 1979 |

| | | | | |
|---|---|---|---|---|
| SRA Survival Skills | 6-Adult | 1976 | SRA | TCB Jul 77 |
| STS Educational Development Series: Scholastic Tests | 2-12 | 1976 | STS | MMY 27 |
| Senior High Assessment of Reading Performance Forms A, B, C (SHARP) | 10-12 | 1978 | CTB | TCB Jul 77 Pg. 11 (Form A) NCME Winter 77 (Form B) |
| Stories about Real-Life Problems | 5-8 | — | NIU | NOT May 79 |
| Test of Consumer Competencies | 8-12 | 1976 | STS | TCB Jan Pg. 6 |
| Test of Everyday Writing Skills (TEWS) | 9-12 | 1978 | CTB | NCME Spring 78 |
| Tests of Performance in Computational Skills (TOPICS) | 9-12 | 1978 | CTB | NCME Winter 77 |
| Wisconsin Test of Adult Basic Education | Adult | — | RFD | FLIT Pg. 48 |

## LANGUAGE ARTS TESTS

| Test | Grade Level(s) | Publication Date | Publisher | Reference |
|---|---|---|---|---|
| Analysis of Skills: Language Arts (ASK: Language Arts) | 2-8 | 1976 | STS | MMY 41 |
| Diagnostic Screening Test: Language | K-Adult | 1977 | SC | NOT Oct 79 |
| Language Arts: IOX Objectives-Based Tests | K-6 | 1974 | IOX | MMY 53 |
| Language Arts: Minnesota High School Achievement Examinations | 7-12 | 1970 | AGS | TIP 90 |
| Writing Test: McGraw-Hill Basic Skills System | 11-12, Adults | 1970 | MHBC | TIP 125 |

# APPENDIX B

## Publishers' Names, Addresses and Telephone Numbers

*This Appendix lists all publishers identified in Appendix A.*

**ACT** The American College Testing Program
P.O. Box 168
Iowa City, Iowa 52240
(319) 356-3711

**AGS** American Guidance Service, Inc.
Publisher's Bldg.
Circle Pines, MN 55014
(612) 786-4343

**ATP** Academic Therapy Publications
28 Commercial Blvd.
Novato, CA 94947
(415) 883-3314

**AW** Addison-Wesley Publishing Co., Inc.
Jacob Way
Reading, MA 01867
(617) 944-3700

**BFA** BFA Educational Media
2211 Michigan Avenue
P.O. Box 1795
Santa Monica, CA 90406
(213) 829-2901

**BMC** Bobbs Merrill Co., Inc.
4300 West 62nd Street
Indianapolis, IN 46268
(317) 298-5400

**CAI** Curriculum Associates, Inc.
5 Esquire Rd.
N. Billerica, MA 01862
(617) 935-8410

**CAL-P** CAL Press, Inc.
76 Madison Ave.
New York, NY 10016
(212) 685-0892

**CARE** The Center for Applied Research in Education, Inc.
Rt. 59
West Nyack, NY 10994
(914) 358-8991

**Croft** Croft Incorporated
4922 Harford Road
Baltimore, MD 21214
(301) 254-5082

**CSDE** California State Dept. of Education
721 Capitol Mall
Sacramento, CA 95814
(916) 445-4688

**CTB** CTB/McGraw Hill
Del Monte Research Park
Monterey, CA 93940
(408) 649-8400

**EDC** Educational Development Corporation
P.O. Box 45663
Tulsa, OK 74145
(91[illegible] 622-4522

**EdITS** [illegible]ITS/Educational and Industrial Testing Service
P.O. Box 7234
San Diego, CA 92107
(714) 222-1666

**ERB** Educational Records Bureau
Educational Testing Service
Box 619
Princeton, NJ 08540
(609) 921-9000

**EPS** Educators Publishing Service
75 Moulton St.
Cambridge, MA 02138
(617) 547-6706

**HAYES** Hayes Educational Laboratory
7040 North Portsmouth Ave.
Portland, OR 97203
(503) 285-3745

**IILC** Imperial International Learning Corp.
Box 548
Kankakee, IL 60901
(815) 933-7735

**IOX** Instructional Objectives Exchange
Box 24095
Los Angeles, CA 90024
(213) 474-4531

**Jastak** Jastak Associates, Inc.
1526 Gilpin Ave.
Wilmington, DE 19806
(302) 652-4990

**McGrath** McGrath Publishing Co.
P.O. Box 9001
Wilmington, NC 28402
(919) 763-3757

**Merrill** Charles E. Merrill Publishing Co.
1300 Alum Creek Drive
Columbus, OH 43216
(614) 258-8441

**MHBC** McGraw Hill Book Co.
1221 Ave. of the Americas
New York, NY 10020
(212) 997-1221

**NCS** NCS Interpretive Scoring Systems
4401 West 76th St.
Minneapolis, MN 55435
(800) 328-6290

**NIU** Northern Illinois University
Alan M. Voelker
Curriculum and Instruction
Dekalb, IL 60115
(815) 753-1000

**NMS** New Mexico State Dept. of Education
Monitor
Education Bldg.
State Capitol
Santa Fe, NM 87501
(505) 827-2429

**Psy. Corp.** The Psychological Corporation
304 E. 45th Street
New York, NY 10017
(212) 888-3500

**RFD** Rural Family Development Program
University Extension
University of Wisconsin
P.O. Box 1379
Madison, WI 53701
(608) 262-1234

**RS** The Riverside Publishing Company
1919 South Highland Avenue
Lombard, IL 60148
(312) 629-9700

**SC** Stoelting Co.
1350 S. Kostner
Chicago, IL 60623
(312) 522-4500

**SRRC** Southwest Regional Resource Center
127 South Franklin
Juneau, AK 99801
(907) 586-6806

**SF** Scott Foresman & Co.
1900 East Lake Ave.
Glenview, IL 60025
(312) 729-3000

**SOI** SOI Institute
214 Main St.
El Segundo, CA 90245
(213) 322-5995

**SRA** Science Research Associates, Inc.
155 N. Wacker Dr.
Chicago, IL 60606
(800) 621-0664

**STS** Scholastic Testing Service, Inc.
480 Meyer Road
Bensenville, IL 60106
(312) 766-7150

**TCP** Teachers College Press
Teachers College
525 West 120th St.
New York, NY 10027
(212) 678-3929

**USDL** United States Dept. of Labor
Bureau of Labor Statistics
1515 Broadway
New York, NY 10036
(212) 399-5405

**Winch** B. L. Winch and Associates
45 Hitching Post Dr.
Rolling Hills Estates, CA 90274
(213) 547-1240

**Zweig** Richard L. Zweig Associates, Inc.
20800 Beach Blvd.
Huntington Beach, CA 92648
(714) 536-8877

# APPENDIX C

## Reference Materials Describing and Reviewing the Tests

### Mental Measurements Yearbook (MMY)

Of all sources cited, *The Eighth Mental Measurement Yearbook* provided the most comprehensive information on tests. To best utilize the source, read the introductory section, "How to Use This Yearbook." Tests are indexed in the yearbook by test number. That number is provided in the previous test lists as the follow-up reference number. However, the number of any test can also be located via the yearbook title or subject indices. Information provided about each test includes:

Title
Description of the groups for which the test is intended
Date of copyright or publication
Acronym
Part scores
Individual or group test
Forms, parts, and levels
Pages
Machine-scorable answer sheets
Costs
Scoring and reporting services
Time
Author
Publisher
Foreign adaptation
Sublisting
Cross references

Additional references to published articles, books and unpublished theses on the construction, validity, use and limitations of each test are reported as part of each test entry. Original reviews of each test by independent measurement experts are provided.

### Tests in Print (TIP)

The companion volume to the *Yearbook* is *Tests in Print II*. It provides the reader with similar but much less detailed information on tests. Again, a section entitled "How to Use This Book" is provided. *Tests in Print II* presents a bibliography of all known tests published for English-speaking subjects and an index to all tests published in previous editions of the *Mental Measurement Yearbook*. TIP II provides the following information:

Title
Test population
Copyright date
Acronym
Special comments
Part scores
Author
Publisher
Foreign adaptations
Cross references within TIP II
Sublistings

### NCME Measurement News (NCME)

The "NCME Measurement News," the official newsletter of the National Council on Measurement in Education, provides a brief description of recently published tests. Information includes publisher, copyright, subject matter, levels, grade, interpreting manuals and costs. It is suggested that individuals desiring additional information contact the publisher directly using the addresses provided. Tests appearing in the newsletter in the newsletter do not represent endorsement by the NCME or its staff.

### News on Tests (NOT)

ETS "News on Tests" and its predecessor, the "Test Collection Bulletin," provide descriptions similar to the NCME newsletter. The test title, author, publisher and address copyright and grade level are included, with a brief statement of content and levels. Information provided is descriptive rather than evaluative, and Educational Testing Service "News on Tests" also includes announcements of new publications relating to testing, conferences, and available test bibliographies.

### Tests of Functional Literacy (FLIT)

The review of currently available *Tests of Functional Adult Literacy* provides information on the characteristics and quality of a series of standardized, criterion referenced, and informal tests of literacy. Included in the descriptive profiles of tests is information on publisher, content and skill coverage, availability of alternate forms, administration procedures, materials needed, scoring procedures, interpretation procedures, validity and reliability.

The reader is urged to take advantage of these informational documents and to contact test publishers for complete information on available tests.